Database, Protocols and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication ⋆

Norman Poh * Samy Bengio

Rue du Simplon 4, CH-1920 Martigny, Switzerland.

IDIAP Research Institute, Rue du Simplon 4, CH-1920, Martigny.

Abstract

Fusing the scores of several biometric systems is a very promising approach to improve the overall system's accuracy. Despite many works in the literature, it is surprising that there is no coordinate d effort in making a benchmark database available. It should be noted that fusion in this context consists not only of multimodal fusion, but also intramodal fusion, i.e., fusing systems using the same biometric modality but different features, or same features but using different classifiers. Building baseline systems from scratch often prevents researchers from putting more efforts in understanding the fusion problem. This paper describes a database of scores taken from experiments carried out on the XM2VTS face and speaker verification database. It then proposes several fusion protocols and provides some state-of-the-art tools to evaluate the fusion performance.

Key words: biometrics authentication, pattern recognition, score-level fusion, benchmark

Authors' Biography

Norman Poh Hoon Thian obtained his M.Sc. degree from the University of Science, Penang, Malaysia and obtained DEA (Diplome d'Étude Approfondie) from Université Louis Pasteur, Strasbourg, France. He is currently a Ph.D. student and research assistant at the IDIAP Research Institute, Switzerland. His research interest is in biometric authentication and various pattern recognition-related algorithms, particularly in fusion based on classifiers and different features.

Samy Bengio is a senior researcher and the machine learning group leader at the IDIAP Research Institute since 1999, where he supervises PhD students and postdoctoral fellows working on many areas of machine learning such as support vector machines, time series prediction, mixture models, large-scale problems, speech and speaker recognition, multi-modal problems, brain computer interfaces, and many more. He has obtained his PhD in computer science from Universite de Montreal (1993), and spent three post-doctoral years at CNET, the research center of France Telecom, and INRS-Telecommunications (Montreal). He then worked as a researcher for CIRANO, an economic and

⋆ This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. The author also thank Julian Fierrez-Aguilar for giving constructive comments. This publication only reflects the authors' view.
* Tel: +41.27.721.77.53, Fax: +41.27.721.12
Email addresses: `norman@idiap.ch` (Norman Poh), `bengio@idiap.ch` (Samy Bengio).
URL: `http://www.idiap.ch` (Norman Poh).

financial academic research center, applying learning algorithms to finance. Before joining IDIAP, he was a research director at Microcell Labs, a private research center in mobile telecommunications. His current interests include all theoretical and applied aspects of learning algorithms.

1 Motivation

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [1]. However, today, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. Biometric data is often noisy because of deformable nature of biometric traits, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

One very promising approach to improve the overall system's accuracy is to fuse the scores of several biometric systems [2]. Despite many works in the literature, e.g. [3,4], it is surprising that there is no coordinated effort in making a benchmark database available for such task. This work is one step towards better sharing of scores to *focus* on better understanding of the fusion mechanism.

In the literature, there are several approaches towards studying fusion. One practice is to used virtual identities whereby a biometric modality from one person is paired with the biometric modality of another person. From the experiment point of view, these biometric modalities belong to the same person. The resultant user is known as a *chimeric user*. While this practice is somewhat accepted in the literature, it was questioned that whether this was a right thing to do or not during the 2003 Workshop on Multimodal User Authentication [5]. The fundamental issue here is the independence assumption that two or more biometric traits of a single person are independent from each other. To the best of our knowledge, there is no work in the literature that approves or disapproves such assumption. A pioneer study [**?**] showed that performance measured with experiments carried on chimeric users do not necessarily reflect the performance with true multimodal users. Another practice is more reasonable: use off-the-shelf biometric systems [6] and quickly acquire scores. While this is definitely a better solution, committing to acquire the systems and to collect the data is admittedly a very time-consuming process. None of the mentioned approaches prevails over the others in understanding the problem of fusion. There are currently on-going but independent projects in the biometric community to acquire multimodal biometric databases. They are listed below [1]

- BANCA [7] – face and speech modalities [2].

[1] This list is by no means exhaustive. An updated list is being maintained at http://www.idiap.ch/∼norman/fusion/main.php?bodyfile=bmark_db.html

[2] http://www.ee.surrey.ac.uk/banca

- XM2VTS [8] – face and speech modalities [3].
- BIOMET [9] – contains face, speech, fingerprint, hand and signature modalities
- MYCT [10] – ten-print fingerprint and signature modalities [4].
- University of Notre Dame Biometrics multimodal databases – face, ear profile and hand modalities acquired using visible, Infrared-Red and range sensors at different angles [5].
- the FRGC database – face modality captured using camera at different angles and range sensors in different controlled or uncontrolled settings [6].

Other benchmarking efforts can be found in [11] [7] involving signature modality and [12] [8] fingerprint modality. As a matter of fact, most reported works in the literature about fusion often concentrates on treatment of the baseline systems. While baseline systems are definitely important, the subject of fusion is unfortunately downplayed. Hence, we propose here not only to publish scores, but also to provide a clear documentation of the baseline systems, well-defined *fusion protocols* and provide a common set of evaluation tools so that experimental results can be compared. The scores are taken from the publicly available XM2VTS face and speech database [9]. It should be mentioned here that there exists another software tool that analyses biometric error rate called PRESS[13] [10]. However, it does not include the DET curve. The tools

[3] http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb

[4] http://turing.ii.uam.es/bbdd_EN.html

[5] (http://www.nd.edu/∼cvrl/UNDBiometricsDatabase.html

[6] http://www.frvt.org/FRGC

[7] http://www.cs.ust.hk/svc2004

[8] http://bias.csr.unibo.it/fvc2004

[9] The database and tools are available in http://www.idiap.ch/∼norman/fusion

[10] Available in. http://it.stlawu.edu/∼msch/biometrics/downloads.htm

proposed here, together with the database, provide a new plot called Expected Performance Curve (EPC) [14] and a significant test specially designed to test the Half Total Error Rate (HTER) [15].

Section 2 explains the XM2VTS database, the Lausanne Protocols and the proposed Fusion Protocols. Section 3 documents the 8 baseline systems that can be used for fusion. Section 4 presents the evaluation criteria, i.e., how experiments should be reported and compared. A set of evaluation tools to facilitate experimentation are presented in Section 5. Some experiments using the proposed fusion protocol are reported in Section 6. This is followed by conclusions in Section 7. Together with the evaluation tool, a recommended cross-validation procedure is also given and is presented in the appendix.

2 Database and Protocols

2.1 The XM2VTS database and the Lausanne Protocols

The XM2VTS database [16] contains synchronised video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence.

The database is divided into three sets: a training set, an evaluation set and a test set. The training set (LP Train) was used to build client models, while the evaluation set (LP Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (LP

Test) was used to estimate the performance.

The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as **LP1** and **LP2** in this paper. In both configurations, the test set remains the same. Their difference is that there are three training shots per client for LP1 and four training shots per client for LP2. Table 1 is the summary of the data. The last column of Table 1 is explained in Section 2.2.

Note that LP Eval's of LP1 and LP2 are used to calculate the optimal thresholds that will be used in LP Test. Results are reported only for the test sets, in order to be as unbiased as possible (using an *a priori* selected threshold). More details can be found in [8].

2.2 The Fusion Protocols

The fusion protocols are built upon the Lausanne Protocols. Before the discussion, it is important to distinguish two categories of approaches: *client-independent* and *client-dependent* fusion approaches. The former approach has only a global fusion function that is common to *all* identities in the database. The latter approach has a different fusion function for a different identity. It has been reported that client-dependent fusion is better than client-independent fusion, given that there are "enough" client-dependent score data. Examples of client-dependent fusion approach are client-dependent threshold [17], client-dependent score normalisation [18] or different weighing of expert opinions using linear [19] or non-linear combination [20]. The fusion protocols that are described here can be client-dependent or client-independent.

It should be noted that one can fuse any of the 8 baseline experiments in LP1 and 5 baseline experiments in LP2 (to be detailed in Section 3). We propose a full combination of all these systems. This protocol is called **FP-full**. Hence, there are altogether $2^8 - 8 - 1 = 248$ possible combinations for LP1 and $2^5 - 5 - 1 = 26$ for LP2. The reasons for minus one and minus the number of experts are that using zero expert and using a single expert are not valid options. However, some constraints are useful. For instance, in some situations, one is constrained to using a single biometric modality. In this case, we propose an intramodal fusion (**FP-intramodal**) . When no constraint is imposed, we propose a full combination (**FP-multimodal**). FP-intramodal contains $2^5-5-1 = 26$ face-expert fusion experiments for LP1, $2^3-3-1 = 4$ speech-expert fusion experiments for LP1, 1 face-expert fusion experiment for LP2 and $2^3 - 3 - 1 = 4$ speech expert-fusion experiments for LP2. Hence, FP-intramodal contains 35 fusion experiments. The second protocol contains $\sum_{m=1}^{5} \sum_{n=1}^{3} ({}^5C_m {}^3C_n) = 217$ combinations, where ${}^nC_k$ is "n choose k". As can be seen, the first three fusion protocols contain an exponential number of combinations. For some specific study, it is also useful to introduce a smaller set of combinations, each time using only two baseline experts, according to the nature of the base-expert. This protocol is called **FP-2**. Three categories of fusion types have been identified under FP-2, namely multimodal fusion (using different biometric traits), intramodal fusion with *different* feature sets and intramodal fusion with the *same* feature set but *different* classifiers. There are altogether 32 such combinations. The pairings of base-experts for fusion are shown in Table 2 for LP1 and LP2.

Note that there are 8 biometric samples in the XM2VTS database on a per client basis. They are used in the following decomposition: 3 samples are used to train the baseline experts in LP1 (and 4 in LP2) on LP Train. There are remaining 3 samples in the in LP1 Eval (and only 2 in LP2 Eval). Finally, for both protocols, 2 client accesses for testing in the *test set*. Because fusion classifiers cannot be trained using scores from the *training set*, or they are simply not available in the current settings, we are effectively using the LP Eval to train the fusion classifiers and then LP Test to test the fusion classifiers' performance on the LP Test. To avoid confusion in terminology used, we call LP Eval as the *fusion development set* and LP Test as the *fusion evaluation set*.

Because we are left with two sets of scores (fusion development and evaluation sets), we propose to use Algorithm 1 detailed in Appendix A, so that we can still obtain two sets of *fused scores*: one on the fusion development set and one on the fusion evaluation set. This is important because a threshold will have to be chosen from the development set and the same threshold will be fixed *a priori* on the evaluation set. In this way, the final reported fusion performance will be unbiased with respect to the threshold parameter.

The use of Algorithm 1 is only highly recommended *but not obligatory*. One can also train a fusion classifier using the evaluation set and also using the same set to output fused scores. Although these scores are biased since they are used to optimise the classifier parameters, we found that, in practice, the threshold calculated this way is often surprisingly acceptable for the test set.

3 Baseline System Description

There are altogether 8 baseline systems [11] . All the 8 baseline systems were used in LP1. On the other hand, 5 out of 8 were used in LP2. This results in 13 baseline experiments (for LP1 and LP2). The following explanation describe these systems in terms of their features, classifiers, and the complete system which is made up of the pair (feature type, classifier).

3.1 Face and Speech Features

The face baseline experts are based on the following features:

(1) **FH**: normalised **f**ace image concatenated with its RGB **H**istogram (thus the abbreviation **FH**) [21].

(2) **DCTs**: DCTmod2 features [22] extracted from face images with a size of 40×32 (rows $\times$ columns) pixels. The Discrete Cosine Transform (DCT) coefficients are calculated from an 8×8 window with horizontal and vertical overlaps of 50%, i.e., 4 pixels in each direction. Neighbouring windows are used to calculate the "delta" features. The result is a set of 35 feature vectors, each having a dimensionality of 18. (**s** indicates the use of this small image compared to the bigger size image with the abbreviation **b**.)

(3) **DCTb**: Similar to DCTs except that the input face image has 80×64 pixels. The result is a set of 221 feature vectors, each having a dimen-

[11] In our website, we also accept public contribution of score files. Each contributor will have to explain clearly their baseline system. Hence, more fusion protocols may be incorporated in the future.

sionality of 18.

The speech baseline experts are based on the following features:

(1) **LFCC**: The Linear Filter-bank Cepstral Coefficient (LFCC) [23] speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients are computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the linear filter-bank. The first temporal derivatives are added to the feature set.

(2) **PAC**: The Phase Auto-Correlation Mel Filter-bank Cepstral Coefficient (PAC-MFCC) features [24] are derived with a window length of 20 milliseconds and each window moves at a rate of 10 milliseconds. 20 DCT coefficients are computed to decorrelate the 30 coefficients obtained from the Mel-scale filter-bank. The first temporal derivatives are added to the feature set.

(3) **SSC**: Spectral Subband Centroid (SSC) features, originally proposed for speech recognition [25], were used for speaker authentication in [26]. It was found that mean-subtraction could improve these features significantly. The mean-subtracted SSCs are obtained from 16 coefficients. The γ parameter, which is a parameter that raises the power spectrum and controls how much influence the centroid, is set to 0.7 [27]. Also, the first temporal derivatives are added to the feature set.

3.2 *Classifiers*

Two different types of classifiers were used for these experiments: Multi-Layer Perceptrons (MLPs) and a Bayes Classifier using Gaussian Mixture Models (GMMs) [28]. While in theory both classifiers could be trained using any of the previously defined feature sets, in practice MLPs are better at matching feature vectors of fixed-size while GMMs are better at matching sequences (feature vectors of unequal size). Whatever the classifier is, the hyper-parameters (e.g. the number of hidden units for MLPs or the number of Gaussian components for GMMs) are tuned on the evaluation set LP1 Eval. The same set of hyper-parameters are used in both LP1 and LP2 configurations of the XM2VTS database.

For each client-specific MLP, the feature vectors associated to the client are treated as positive patterns while all other feature vectors *not* associated to the client are treated as negative patterns. All MLPs reported here were trained using the stochastic version of the error-back-propagation training algorithm [28].

For the GMMs, two competing models are often needed: a world and a client-dependent model. Initially, a world model is first trained from an external database (or a sufficiently large data set) using the standard Expectation-Maximisation algorithm [28]. The world model is then adapted for each client to the corresponding client data using the Maximum-A-Posteriori adaptation [29] algorithm.

The baseline experiments based on DCTmod2 feature extraction were reported in [30] while those based on normalised face images and RGB histograms (FH features) were reported in [21]. Details of the experiments, coded in the pair **(feature, classifier)**, for the face experts, are as follows:

(1) **(FH, MLP)** Features are normalised **F**ace concatenated with **H**istogram features. The client-dependent classifier used is an MLP with 20 hidden units. The MLP is trained with geometrically transformed images [21].

(2) **(DCTs, GMM)** The face features are the DCTmod2 features calculated from an input face image of 40×32 pixels, hence, resulting in a sequence of 35 feature vectors each having 18 dimensions. There are 64 Gaussian components in the GMM. The world model is trained using *all the clients* in the training set [30].

(3) **(DCTb, GMM)** Similar to (DCTs,GMM), except that the features used are DCTmod2 features calculated from an input face image of 80×64 pixels. This produces in a sequence of 221 feature vectors each having 18 dimensions. The corresponding GMM has 512 Gaussian components [30].

(4) **(DCTs, MLP)** Features are the same as those in (DCTs,GMM) except that an MLP is used in place of a GMM. The MLP has 32 hidden units [30]. Note that in this case a training example consists of a *big single* feature vector with a dimensionality of 35×18. This is done by simply concatenating 35 feature vectors each having 18 dimensions [12].

[12] This may explain why MLP, an inherently discriminative classifier, has worse performance compared to GMM, a generative classifier. With high dimensionality yet having only a few training examples, the MLP cannot be trained optimally.

(5) **(DCTb, MLP)** The features are the same as those in (DCTb,GMM) except that an MLP with 128 hidden units is used. Note that in this case the MLP in trained on a *single* feature vector with a dimensionality of 221×18 [30].

and for the speech experts:

(1) **(LFCC, GMM)** This is the Linear Filter-bank Cepstral Coefficients (LFCC) obtained from the speech data of the XM2VTS database. The GMM has 200 Gaussian components, with the minimum relative variance of each Gaussian fixed to 0.5, and the MAP adaptation weight equals 0.1. This is the best known model currently available [31] under clean conditions.

(2) **(PAC, GMM)** The same GMM configuration as in LFCC is used. Note that in general, 200-300 Gaussian components would give about 1% of difference of HTER [31]. This system is particularly robust to very noisy conditions (less than 6 dBs, as tested on the NIST2001 one-speaker detection task).

(3) **(SSC, GMM)** The same GMM configuration as in LFCC is used [27]. This system is known to provide an optimal performance under moderately noisy conditions (18-12 dBs, as tested on NIST2001 one-speaker detection task).

This may affect its generalisation on unseen examples. By treating the features as a sequence, GMM was able to generalise better and hence is more adapted to this feature set.

4 Evaluation Criteria

There are three important concepts about evaluation of a biometric system: types of errors in biometric authentication, threshold criterion and evaluation criterion. The types of errors are false acceptance and false rejection (see Section 4.1). A *threshold criterion* refers to a strategy to choose a threshold which is necessarily tuned on *a development set*. An *evaluation criterion* is used to measure the final performance and is necessarily calculated on *an evaluation set*. Both are discussed in Sections 4.2 and 4.3. Section 4.4 addresses the issue of measuring the gain due to fusion as compared to individual baseline expert systems. Section 4.5 discusses how to visualise the evaluation criterion. It also deals with the case of visualising the performance of several systems using one single curve. Finally, Section 4.6 presents a significance test that can be used to compare the performance of *two* systems using the evaluation criterion.

4.1 Types of Errors

A fully operational biometric system makes a decision using the following *decision function*:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y(\mathbf{x}) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{1}$$

where $y(\mathbf{x})$ is the output of the underlying expert supporting the hypothesis that the biometric sample received $\mathbf{x}$ belongs to a client. The variables that follow will be derived from $y(\mathbf{x})$. For simplicity, we write y instead of $y(\mathbf{x})$. The same convention applies to variables that follow. Because of the accept-reject outcomes, the system may make two types of errors, i.e., false acceptance (FA) and false rejection (FR). Normalised versions of FA and FR are often used and called false acceptance rate (FAR) and false rejection rate (FRR), respectively. They are defined as:

$$\mathrm{FAR}(\Delta) = \frac{\mathrm{FA}(\Delta)}{NI}, \tag{2}$$

$$\mathrm{FRR}(\Delta) = \frac{\mathrm{FR}(\Delta)}{NC}. \tag{3}$$

where FA and FR count the number of FA and FR accesses, respectively; and NI and NC are the total number of impostor and client accesses, respectively.

4.2 Threshold Criterion

To choose an "optimal threshold" Δ, it is necessary to define a threshold criterion. This has to be done on a development set. Two commonly used criteria are the Weighted Error Rate (WER) and Equal Error Rate (EER). WER is defined as:

$$\mathrm{WER}(\alpha, \Delta) = \alpha \mathrm{FAR}(\Delta) + (1 - \alpha)\ \mathrm{FRR}(\Delta), \tag{4}$$

where $\alpha \in [0, 1]$ balances between FAR and FRR. A special case of WER is EER, which assumes that the costs of FA and FR are equal. It further assumes that the class prior distributions of client and impostor accesses are equal. As a result $\alpha = 0.5$. In this case, Eqn. (4) becomes:

$$\mathrm{EER}(\Delta) = \frac{1}{2}\left(\mathrm{FAR}(\Delta) + \ \mathrm{FRR}(\Delta)\right). \tag{5}$$

Let Δ^*_α be the optimal threshold that *minimises* WER on a *development set*. It can be calculated as follows:

$$\Delta^* = \arg\min_{\Delta} |\alpha \mathrm{FAR}(\Delta) - (1-\alpha)\ \mathrm{FRR}(\Delta)|. \tag{6}$$

Note that one could have also used a second minimisation criterion:

$$\Delta^* = \arg\min_{\Delta} \mathrm{WER}(\alpha, \Delta). \tag{7}$$

In theory, these two minimisation criteria should give identical results. In practice however, they do not, because FAR and FRR are empirical functions and are not smooth. Eqn. (6) ensures that the difference between weighted FAR and weighted FRR are as small as possible while Eqn. (7) ensures that the sum of the two weighted terms are minimised. Because FAR is a decreasing function while FRR is an increasing function of threshold, the first criterion takes advantage of this additional information while the second criterion does not. Hence, the first criterion can more accurately estimate the threshold and is used for evaluation in this study. Note that the EER criterion can be calculated similarly by fixing $\alpha = 0.5$.

4.3 Evaluation Criterion

Having chosen an optimal threshold using the WER threshold criterion discussed previously, the final performance is measured using Half Total Error Rate (HTER). Note that the threshold is found with respect to a given α. It is defined as:

$$\mathrm{HTER}(\Delta^*_\alpha) = \frac{\mathrm{FAR}(\Delta^*_\alpha) +\ \mathrm{FRR}(\Delta^*_\alpha)}{2}. \tag{8}$$

It is important to note that the FAR and FRR do not have the same *resolution*. Because there are more simulated impostor accesses than the client accesses, FRR changes more drastically when falsely rejecting a client access whereas FAR changes less drastically when falsely accepting an impostor access. Hence, when comparing the performance using HTER(Δ^*_α) from two systems (at the *same* Δ^*_α), the question of whether the HTER difference is significant or not has to take into account the imbalanced numbers of client and impostor accesses. This is discussed in Section 4.6.

Finally, it is important to note that HTER in Eqn. (8) is identical to EER in Eqn. (5) except that HTER is a *performance measure* (calculated on an *evaluation set* whereas EER is a *threshold criterion* optimised on a *development set*. Because of their usage in different context, EER should not be interpreted as a performance measure (in place of HTER) to compare the performance of different systems. Such practice, to our opinion, leads to an *unrealistic* comparison. The argument is that in an actual operating system, the threshold has to be fixed *a priori*. This subject is further discussed in Section 4.5. To distinguish these two concepts, when discussing HTER calculated on a development set using a threshold criterion also calculated on the same set, the HTER should be called *a posteriori* HTER. When discussing HTER calculated on an evaluation set with a threshold optimised on a development set, the HTER should be called *a priori* HTER.

4.4 Measuring the Gain

This section presents the "gain ratio" to answer the question: "how much gain can one obtain out of a given fusion experiment as compared to the baseline systems?". Suppose that there are $i = 1, \ldots, N$ baseline expert systems. HTER_i is the HTER evaluation criterion (measured on an *evaluation* set) associated to expert i and HTER_{COM} is the HTER associated to the combined system. The "gain ratio" β has two definitions, as follow:

$$\beta_{mean} = \frac{\text{mean}_i(\text{HTER}_i)}{\text{HTER}_{COM}} \tag{9}$$

$$\beta_{min} = \frac{\text{min}_i(\text{HTER}_i)}{\text{HTER}_{COM}}, \tag{10}$$

where β_{mean} and β_{min} are the proportion of the HTER of the combined (fused) expert with respect to the mean and the minimum HTER of the underlying experts $i = 1, \ldots, N$. According to our previous work, it is found theoretically and empirically that $\beta_{mean} \geq 1$ [32]. β_{min}, on the other hand, is a more realistic criterion, i.e., one wishes to obtain better performance than the best underlying expert, but there is no analytical proof that $\beta_{min} \geq 1$.

4.5 Visualising the Performance

Perhaps the most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [33], which is actually a Receiver Operator Curve (ROC) curve plotted on a non-linear scale. It has been pointed out [14] that two DET curves resulted from two systems are not comparable because such comparison does not take into account how the thresholds are selected. In fact, this holds down to compare two DET curves at a given common threshold chosen *a posteriori*. It was argued [14] that such threshold should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [14] was proposed. We will adopt this evaluation method, which is also in coherence with the original Lausanne Protocols defined for the XM2VTS database.

The EPC curve simply plots HTER (in Eqn. (8)) versus α (as found in Eqn. (4)), since different values of α give rise to different values of HTERs. The EPC curve can be interpreted in the same manner as the DET curve, i.e., the lower the curve is, the better the performance but for the EPC curve, the comparison is done at a given cost (controlled by α). An example of EPC Curve is shown in Figure 1(a).

One advantage of EPC curve is that it can plot a pooled curve from several experiments (although this could also be done with DET curves, but in this case, it would not take the threshold selection into account). For instance, to compare two methods over M experiments, only one pooled curve is necessary. This is done by calculating HTER at a given α point by taking into account all the false acceptance and false rejection accesses over all M experiments. The pooled FAR and FRR across $j = 1, \ldots, M$ experiments for a given $\alpha \in [0, 1]$ is defined as follow:

$$\mathrm{FAR}^{pooled}(\Delta^*_\alpha) = \frac{\sum_{j=1}^{M} \mathrm{FA}(\Delta^*_\alpha(j))}{NI \times M}, \tag{11}$$

and

$$\mathrm{FRR}^{pooled}(\Delta^*_\alpha) = \frac{\sum_{j=1}^{M} \mathrm{FR}(\Delta^*_\alpha(j))}{NC \times M}, \tag{12}$$

where $\Delta^*_\alpha(j)$ is the optimised threshold at a given α for expert j, NI is the number of impostor accesses and NC is the number of client accesses. FA and FR count the number of false acceptance and the number of false rejection at a given threshold $\Delta^*_\alpha(j)$. The pooled HTER is defined similarly as in Eqn. (8)

by using the pooled versions of FAR and FRR.

4.6 HTER Significance Test

Although there exists several statistical significance tests in the literature such as McNemar's Test [34] and *Asymptotic Performance*, it has been shown that HTER significance test [15] better reflects the imbalance nature of precision in FAR and FRR.

When comparing two EPC curves that are very close to each other, with possible overlaps, it is interesting to know if the difference between the two HTERs, at any given point of α, is significant or not. In this case, it is recommended to employ a two-sided significance test as proposed in [15]. Under some reasonable assumptions, it has been shown [15] that the difference of HTER of two systems (say A and B) is normally distributed with the following variance:

$$\sigma^2_{\text{HTER}} = \frac{\frac{\text{FAR}_A(1-\text{FAR}_A)+\text{FAR}_B(1-\text{FAR}_B)}{4\cdot\text{NI}} + }{\frac{\text{FRR}_A(1-\text{FRR}_A)+\text{FRR}_B(1-\text{FRR}_B)}{4\cdot\text{NC}}}, \tag{13}$$

where HTER_A, FAR_A and FRR_A are HTER, FAR and FRR of the first system labeled A and similarly for the second system labeled B. NI and NC are the total number of impostor accesses and client accesses, respectively. One can then compute the following z-statistics:

$$z = \frac{\text{HTER}_A - \text{HTER}_B}{\sigma_{\text{HTER}}}. \tag{14}$$

Let us define $D(z)$ as the cumulative density of a normal distribution. The significance of z is calculated as $D^{-1}(z)$. In a standard two-sided test, $|z|$ is used. In Eqn. (14), the sign of z is retained so that $z > 0$ implies that $\text{HTER}_A > \text{HTER}_B$. Consequently, $D^{-1}(z) > 0.5$ and vice-versa for $z < 0$. An example of plot of significance test is shown in Figure 1(b). This significance test is performed on EPC curves shown in Figure 1(a). (DCTs,GMM) is system A whereas (PAC,GMM) is system B. Whenever the EPC curve of system B is lower than that of system A (B is better than A), the corresponding significance curve is more than 50%. Below 10% of confidence (or above 90% of confidence) indicates that system B is significantly worse than A (or system A is significantly worse than B).

5 Evaluation Tools

There are several (software) tools provided:

- **The `main` program.** This program handles loading and fusion of all the 32 fusion experiments in FP-2. There is also a sample program showing how the fusion can be done using the mean operator. The output of the `main` program is shown in Table 2.
- **Visualisation tool.** This program plots two-dimensional scatter plot of scores with Gaussian fittings for each class of scores. An example of scatter plot is shown in Figure 2. Gaussian fittings can be very useful to predict fusion [35] especially when the Gaussian hypothesis is true.
- **Score diagnostic tool.** This program takes as input fused scores and calculate HTERs, plots class-dependent distributions, FAR and FRR versus threshold and DET curve. An example of output of this program is shown in Figure 3.

- **EPC curves.** Last but not least, there is also a program that plots Expected Performance Curve or EPC curve [14] (see Figure 1(a)). A detailed explanation of EPC can be found in Section 4.5. Furthermore, another accompanying program actually compares two EPC curves to see if the difference of two systems under comparison is significant or not using a two-sided test [15], at various operating costs (see Figure 1(b)). A detailed explanation of this test can be found in Sec 4.6.

6 Example of Fusion Experiments

Some experiments have been carried out in client-independent fusion setting (e.g. [32]) and client-dependent fusion setting (e.g. [**?**]). In [32] client-independent FP-2 was performed using the mean operator, Multi-Layer Perceptrons, Support Vector Machines [32]. Here, we show only fusion experiments using the mean operator (see Table 2). For multimodal fusion, there are a total of 21 experiments from both LP1 and LP2 protocols. Similarly for fusion with different feature sets (same modality), there are a total of 9 experiments (6 from LP1 and 3 from LP2). Finally, for fusion with different classifiers (same feature set), there are only 2 experiments.

7 Conclusions

In this study, we presented a database, several fusion protocols in different scenarios and a set of evaluation tools to encourage researchers to focus on the problem of biometric authentication score-level fusion. To the best of our knowledge, there has been no work in the literature that provides a benchmark database for score-level fusion. Several practical and state-of-the-art tools are also provided so that experiments can be compared in a realistic and unbiased way.

A Cross-Validation Procedure

Algorithm 1 [36] shows how K-fold cross-validation can be used to estimate the correct value of the hyper-parameters of our fusion model, as well as the decision threshold used in the case of authentication. The basic framework of the algorithm is as follows: first perform K-fold cross-validation on the training set by varying the hyper-parameter, and for each hyper-parameter, select the corresponding decision threshold that minimises Half Total Error Rate (HTER); then choose the best hyper-parameter according to this criterion and perform normal training with the best hyper-parameter on the whole training set; finally test the resultant classifier on the test set with HTER evaluated on the previously found decision threshold.

There are several points to note concerning Algorithm 1: $\mathcal{Z}$ is a set of labeled examples of the form $(\mathcal{X}, \mathcal{Y})$, where the first term is a set of patterns and the second term is a set of corresponding labels. The "train" function receives a hyper-parameter θ and a training set, and outputs an optimal classifier $\hat{F}$ by minimising the HTER on the training set. The "test" function receives a classifier $\hat{F}$ and a set of examples, and outputs a set of scores for each associated example. The "thrd_{HTER}" function returns a *decision threshold* that minimises HTER by minimising $|\mathrm{FAR}(\Delta) - \mathrm{FRR}(\Delta)|$ with respect to the threshold Δ ($\mathrm{FAR}(\Delta)$ and $\mathrm{FRR}(\Delta)$ are false acceptance and false rejection rates, as a function of Δ) while L_{HTER} returns the HTER *value* for a particular

Algorithm 1 Risk Estimation $(\Theta, K, \mathcal{Z}^{train}, \mathcal{Z}^{test})$

REM: Risk Estimation with K-fold Validation. See [36].

Θ : a set of values for a given hyper-parameter

$\mathcal{Z}^i$: a tuple $(\mathcal{X}^i, \mathcal{Y}^i)$, for $i \in \{train, test\}$ where

$\mathcal{X}$: a set of patterns. Each pattern contains scores/hypothesis from base experts

$\mathcal{Y}$: a set of labels $\in \{client, impostor\}$

Let $\cup_{k=1}^{K} \mathcal{Z}^k = \mathcal{Z}^{train}$ and $\mathcal{Z}^i \cap \mathcal{Z}^j = \emptyset \forall_{i,j}$

for each hyper-parameter $\theta \in \Theta$ **do**

  for each $k = 1, \ldots, K$ **do**

    $\hat{F}_\theta = \text{train}(\theta,\, \cup_{j=1, j\neq k}^{K} \mathcal{Z}^j)$

    $\hat{\mathcal{Y}}_\theta^k = \text{test}(\hat{F}_\theta,\, \mathcal{X}^k)$

  end for

  $\Delta_\theta = \text{thrd}_{HTER}\left(\{\hat{\mathcal{Y}}_\theta^k\}_{k=1}^K, \{\mathcal{Y}^k\}_{k=1}^K\right)$

end for

$\theta^* = \arg\min_\theta \left(L_{HTER}\left(\Delta_\theta, \{\hat{\mathcal{Y}}_\theta^k\}_{k=1}^K, \{\mathcal{Y}^k\}_{k=1}^K\right)\right)$

$\hat{F}_{\theta^*} = \text{train}(\theta^*,\, \mathcal{Z}^{train})$

$\hat{\mathcal{Y}}_{\theta^*}^{test} = \text{test}(\hat{F}_{\theta^*},\, \mathcal{X}^{test})$

return $L_{HTER}(\Delta_{\theta^*}, \hat{\mathcal{Y}}_{\theta^*}^{test}, \mathcal{Y}^{test})$

decision threshold. What makes this cross-validation different from classical cross-validation is that there is only one single decision threshold and the corresponding HTER value for all the held-out folds and for a given hyper-parameter θ. This is because it is logical to union scores of all held-out folds into one single set of scores to select the decision threshold (and obtain the corresponding HTER).

References

[1] A. Jain, R. Bolle, S. Pankanti, Biometrics: Person Identification in a Networked Society, Kluwer Publications, 1999.

[2] J. Kittler, G. Matas, K. Jonsson, M. Sanchez, Combining Evidence in Personal Identity Verification Systems, Pattern Recognition Letters 18 (9) (1997) 845–852.

[3] J. Kittler, K. Messer, J. Czyz, Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems, in: Proc. Cost 275 Workshop, Rome, 2002, pp. 17–24.

[4] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, J. Gonzalez-Rodriguez, A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification, in: LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003), Guildford, 2003, pp. 830–837.

[5] J.-L. Dugelay, J.-C. Junqua, K. Rose, M. Turk, Workshop on Multimodal User Authentication (MMUA 2003), no publisher, Santa Barbara, CA, 11–12 December, 2003.

[6] M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain, Multimodal Biometric Authentication Methods: A COTS Approach, in: Workshop on Multimodal User Authentication (MMUA 2003), Santa Barbara, 2003, pp. 99–106.

[7] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, J.-P. Thiran, The BANCA Database and Evaluation Protocol, in: LNCS 2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03, Springer-Verlag, 2003.

[8] J. Lüttin, Evaluation Protocol for the XM2FDB Database (Lausanne Protocol), Communication 98-05, IDIAP, Martigny, Switzerland (1998).

[9] S. Carcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. L. les Jardins, J. Lunter, Y. Ni, D. Petrovska-Delacrtaz, BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities, in: LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03), Guildford, 2003, pp. 845–853.

[10] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, Q.-I. Moro, Biometric on the Internet MCYT Baseline Corpus: a Bimodal Biometric Database, IEE Proc. Visual Image Signal Processing 150 (6) (2003) 395–401.

[11] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, G. Rigoll, Signature Verification Competition 2004, in: LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA), Hong Kong, 2004, pp. 16–22.

[12] D. Maio, D. Maltoni, R. Cappelli, J. Wayman, A. Jain, Fingerprint Verification Competition 2004, in: LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA), Hong Kong, 2004, pp. 1–7.

[13] M. E. Schuckers, C. J. Knickerbocker, Documentation for Program for Rate Estimation and Statistical Summaries PRESS, Department of Mathematics, Computer Science and Statistics St Lawrence University, Canton, NY 13617 and Center for Identification Technology Research, West Virginia University.

[14] S. Bengio, J. Mariéthoz, The Expected Performance Curve: a New Assessment Measure for Person Authentication, in: The Speaker and Language Recognition Workshop (Odyssey), Toledo, 2004, pp. 279–284.

[15] S. Bengio, J. Mariéthoz, A Statistical Significance Test for Person Authentication, in: The Speaker and Language Recognition Workshop

(Odyssey), Toledo, 2004, pp. 237–244.

[16] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Comparison of Face Verification Results on the XM2VTS Database, in: Proc. 15th Int'l Conf. Pattern Recognition, Vol. 4, Barcelona, 2000, pp. 858–863.

[17] J. Saeta, J. Hernando, On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation, in: The Speaker and Language Recognition Workshop (Odyssey), Toledo, 2004, pp. 215–218.

[18] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Target Dependent Score Normalisation Techniques and Their Application to Signature Verification, in: LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA), Hong Kong, 2004, pp. 498–504.

[19] A. Jain, A. Ross, Learning User-Specific Parameters in Multibiometric System, in: Proc. Int'l Conf. of Image Processing (ICIP 2002), New York, 2002, pp. 57–70.

[20] A. Kumar, D. Zhang, Integrating Palmprint with Face for User Authentication, in: Workshop on Multimodal User Authentication (MMUA 2003), Santa Barbara, 2003, pp. 107–112.

[21] S. Marcel, S. Bengio, Improving Face Verification Using Skin Color Information, in: Proc. 16th Int. Conf. on Pattern Recognition, Quebec, 2002, p. unknown.

[22] C. Sanderson, K. Paliwal, Fast Features for Face Authentication Under Illumination Direction Changes, Pattern Recognition Letters 24 (14) (2003) 2409–2419.

[23] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Oxford University Press, 1993.

[24] S. Ikbal, H. Misra, H. Bourlard, Phase Auto-Correlation (PAC) derived Robust Speech Features, in: Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, 2003, pp. 133–136.

[25] K. K. Paliwal, Spectral Subband Centroids Features for Speech Recognition, in: Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Vol. 2, Seattle, 1998, pp. 617–620.

[26] N. Poh, C. Sanderson, S. Bengio, An Investigation of Spectral Subband Centroids For Speaker Authentication, in: LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA), Hong Kong, 2004, pp. 631–639.

[27] N. Poh, C. Sanderson, S. Bengio, An Investigation of Spectral Subband Centroids For Speaker Authentication, Research Report 03-62, IDIAP, Martigny, Switzerland (2003).

[28] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1999.

[29] J. Gauvain, C.-H. Lee, Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains, IEEE Tran. Speech Audio Processing 2 (1994) 290–298.

[30] F. Cardinaux, C. Sanderson, S. Marcel, Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS, in: LNCS 2688, 4th Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03), Guildford, 2003, pp. 911–920.

[31] N. Poh, S. Bengio, Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication, in: The Speaker and Language Recognition Workshop (Odyssey), Toledo, 2004, pp. 199–206.

[32] N. Poh, S. Bengio, Non-Linear Variance Reduction Techniques in Biometric Authentication, in: Workshop on Multimodal User Authentication (MMUA

2003), Santa Barbara, 2003, pp. 123–130.

[33] A. Martin, G. Doddington, T. Kamm, M. Ordowsk, M. Przybocki, The DET Curve in Assessment of Detection Task Performance, in: Proc. Eurospeech'97, Rhodes, 1997, pp. 1895–1898.

[34] T. G. Dietterich, Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms, Neural Computation 10 (7) (1998) 1895–1923.

[35] N. Poh, S. Bengio, Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Task, in: LNCS 3361, Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI, 2004, pp. 159–172.

[36] S. Bengio, C. Marcel, S. Marcel, J. Mariéthoz, Confidence Measures for Multimodal Identity Verification, Information Fusion 3 (4) (2002) 267–276.

Table 1

The Lausanne Protocols of XM2VTS database. The last column shows the terms used in the fusion protocols presented in Section 2.2. LP Eval corresponds to the Fusion protocols' development set while LP Test corresponds to the Fusion Protocols' evaluation set.

Data sets	Lausanne Protocols		Fusion Protocols
	LP1	LP2	
LP Train client accesses	3	4	NIL
LP Eval client accesses	600 (3×200)	400 (2×200)	Fusion dev
LP Eval impostor accesses	40,000 ($25 \times 8 \times 200$)		Fusion dev
LP Test client accesses	400 (2×200)		Fusion eva
LP Test impostor accesses	112,000 ($70 \times 8 \times 200$)		Fusion eva

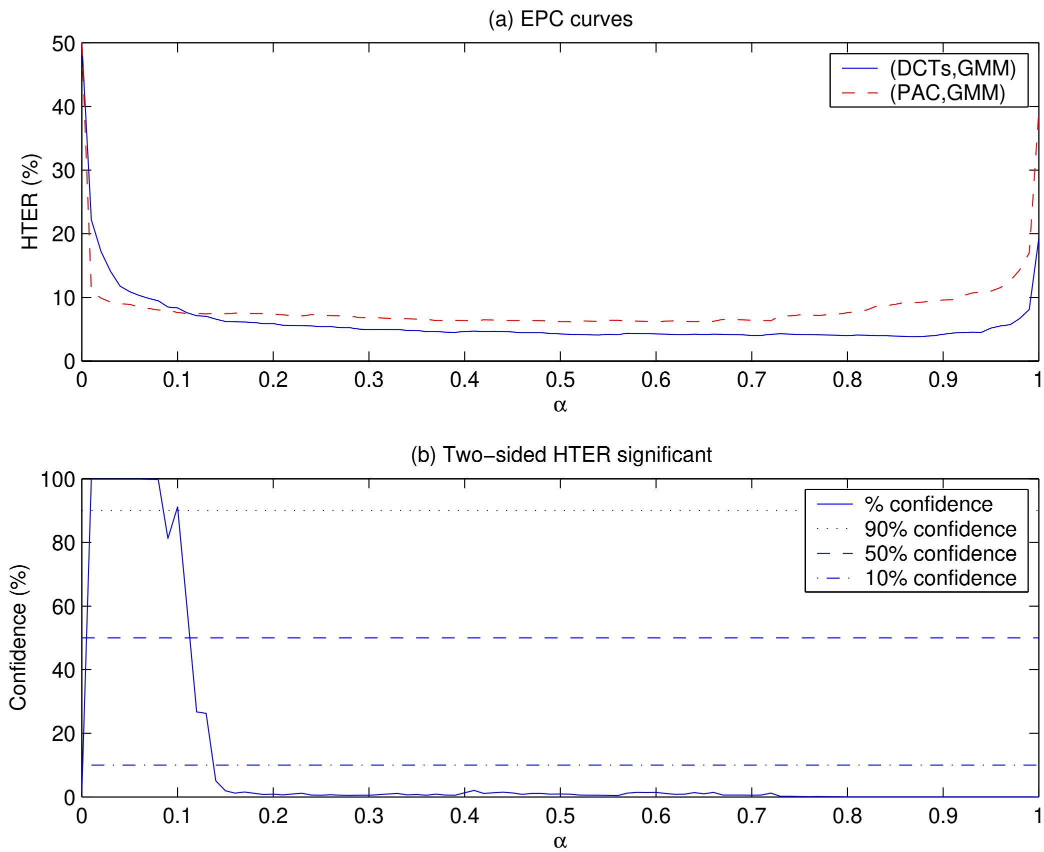

Fig. 1. (a): Expected Performance Curves (EPCs) of two experiments: one is a face system (DCTs,GMM) and the other is speech system (PAC,GMM). (b) HTER significance test of the two EPC curves. Confidence more than 50% implies that the speech expert is better and vice-versa. This is a two-tailed test so two HTERs of a given *cost* α are considered significantly different when the level of confidence is below 10% or above 90% (for a significance level of 20%, in this case for illustration).

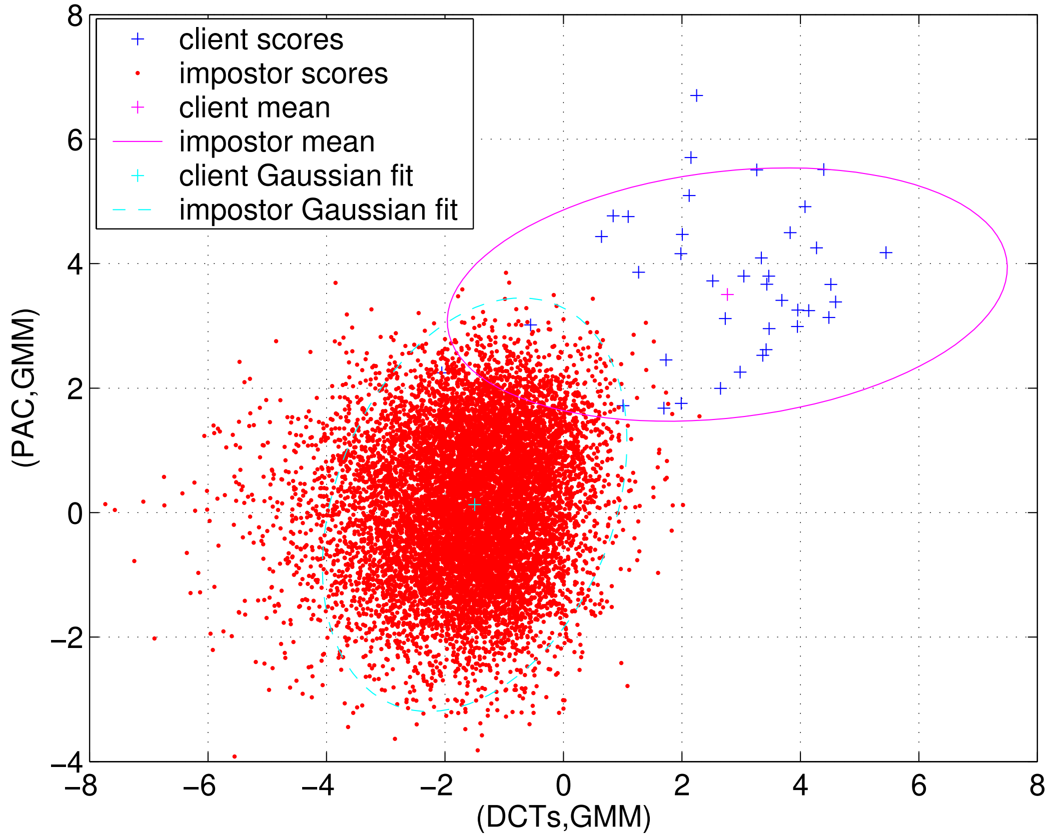

Fig. 2. Scatter plot of two experts. Only 1% of data points are plotted here. Circles overlaid on each cluster of scores is a Gaussian approximation of the respective class (client or impostor).

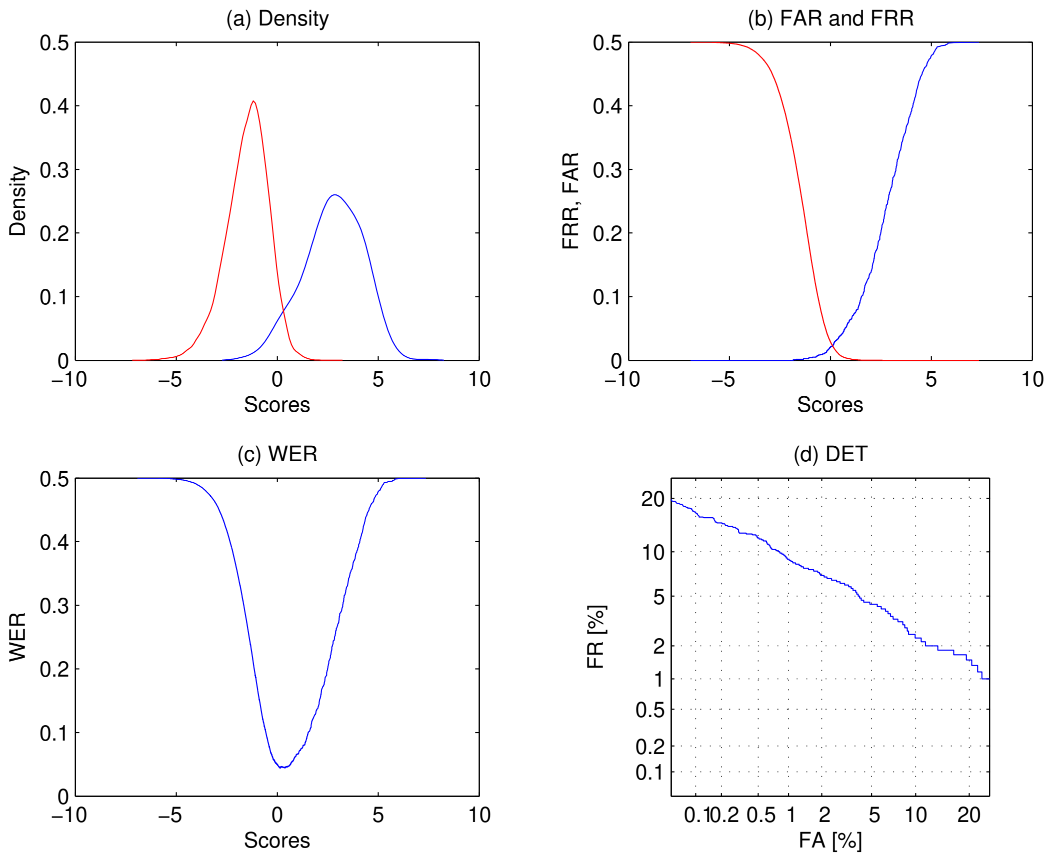

Fig. 3. A score diagnostic plot of two experts. (a): Probabilistic density function of client (right, darker curve) and impostor classs (left, lighter curve). (b): False Rejection Rate (FRR, darker curve) and False Acceptance Rate (FAR, lighter curve) versus the scores (different threshold positions). (c): Weighted Error Rate with $\alpha = 0.5$. (d): Decision Error Trade-off (DET) curve.

Table 2

Results of combining two baseline experts using the mean operator according to FP-2.

(a) Fusion with different modalities for LP1

No.	Fusion candidates	HTER
1	((FH,MLP)(LFCC,GMM))	0.782
2	((FH,MLP)(PAC,GMM))	1.120
3	((FH,MLP)(SSC,GMM))	0.871
4	((DCTs,GMM)(LFCC,GMM))	0.543
5	((DCTs,GMM)(PAC,GMM))	1.436
6	((DCTs,GMM)(SSC,GMM))	1.149
7	((DCTb,GMM)(LFCC,GMM))	0.511
8	((DCTb,GMM)(PAC,GMM))	1.021
9	((DCTb,GMM)(SSC,GMM))	0.752
10	((DCTs,MLP)(LFCC,GMM))	0.840
11	((DCTs,MLP)(PAC,GMM))	1.138
12	((DCTs,MLP)(SSC,GMM))	1.333
13	((DCTb,MLP)(LFCC,GMM))	1.523
14	((DCTb,MLP)(PAC,GMM))	3.664
15	((DCTb,MLP)(SSC,GMM))	3.108

(b) Fusion with different feature sets for LP1

No.	Fusion candidates	HTER
1	((FH,MLP)(DCTs,GMM))	1.280
2	((FH,MLP)(DCTb,GMM))	1.122
3	((FH,MLP)(DCTs,MLP))	1.513
4	((FH,MLP)(DCTb,MLP))	1.960
5	((LFCC,GMM)(SSC,GMM))	1.595
6	((PAC,GMM)(SSC,GMM))	4.225

(c) Fusion with different classifiers for LP1

No.	Fusion candidates	HTER
1	((DCTs,GMM)(DCTs,MLP))	2.388
2	((DCTb,GMM)(DCTb,MLP))	3.063

(d) Fusion with different modalities for LP2

No.	Fusion candidates	HTER
1	((FH,MLP)(LFCC,GMM))	1.122
2	((FH,MLP)(PAC,GMM))	1.513
3	((FH,MLP)(SSC,GMM))	1.960
4	((DCTb,GMM)(LFCC,GMM))	1.836
5	((DCTb,GMM)(PAC,GMM))	2.388
6	((DCTb,GMM)(SSC,GMM))	3.672

(e) Fusion with different feature sets for LP2

No.	Fusion candidates	HTER
1	((FH,MLP)(DCTb,GMM))	1.280
2	((LFCC,GMM)(SSC,GMM))	3.063
3	((PAC,GMM)(SSC,GMM))	2.934